Journal of Information and Technology

Prediction of The Likelihood of Policy Lapsation Using Machine Learning Models: A Case Study of a Life Insurance Company Operating in Kenya

Mogusu Doreen Kerubo, Ngesa Oscar & Kombo Abdallah

ISSN: 2617-3573

Stratford Peer Reviewed Journals and Book Publishing
Journal of Information and Technology
Volume 9||Issue 1||Page 105-125 ||December||2025|
Email: info@stratfordjournals.org ISSN: 2617-3573

Prediction of The Likelihood of Policy Lapsation Using Machine Learning Models: A Case Study of a Life Insurance Company Operating in Kenya

Mogusu Doreen Kerubo[1*], Ngesa Oscar [1] and Kombo Abdallah[1]
***Corresponding Author: kerdee37@gmail.com**
[1].Department of Mathematics, Statistics and Physical Sciences
Taita Taveta University

***How to cite this article*: Mogusu D, K., Ngesa O. & Kombo A. (2025). Prediction of The Likelihood of Policy Lapsation Using Machine Learning Models: A Case Study of a Life Insurance Company Operating in Kenya. *Journal of Information and Technology. Vol* 9(1) pp. 105-125.** https://doi.org/10.53819/81018102t2535

Abstract

Policy lapsation, defined as the cessation of premium payments by policyholders resulting in termination of coverage, poses significant challenges to insurance companies in terms of revenue loss and customer retention. Lapses influence the profitability and liquidity of insurance companies through acquisition cost, and loss of income from renewal premiums; hence needs to be controlled and managed carefully. Leveraging a case study approach, this research explored the effectiveness of various machine learning algorithms in forecasting policy lapsation rates based on historical data and relevant policyholder attributes. Secondary data was obtained from a life insurance company operating in Kenya over the period 2018 to 2023 with 21,891 policyholders. Five classification models (Logistic Regression, Artificial Neural Networks (ANN), Random Forest, XGBoost, and AdaBoost) were trained and evaluated using comprehensive metrics including ROC-AUC, precision-recall AUC, sensitivity, specificity, and accuracy. The results show the strong prediction ability of ensemble models (Random Forest and XGBoost) and identified occupation type, sum assured and payment methods as critical predictors of lapsation. The best overall classifier is Random Forest with an accuracy of 80.6%, precision-recall AUC of 91.2%, and ROC-AUC of 88.2% with balanced specificity (80.1%) and sensitivity (81.1%). XGBoost showed a ROC-AUC of 87.5% and accuracy of 80.3%. The findings underscore the efficacy of ensemble methods, particularly Random Forest, in predicting lapsation risks, offering insurers actionable insights to proactively manage customer retention. This study contributes to the body of knowledge on actuarial analytics by validating machine learning applications in lapse prediction and provides a framework for implementing data-driven decision-making in insurance risk management.

Keywords: *Life Insurance, Policy Lapse, Machine Learning*

Stratford Peer Reviewed Journals and Book Publishing
Journal of Information and Technology
Volume 9||Issue 1||Page 105-125 ||December||2025|
Email: info@stratfordjournals.org ISSN: 2617-3573

1. Introduction

A salient characteristic of property, and human life is the intrinsic risk posed by various activities and actions that surround our environment (Mishr, 2016). For example, human life is susceptible to all forms of threats, such as health pandemics, climate driven natural disasters, human driven disasters including violent conflict and accident (Dorfman, 1998). These unforeseen circumstances can interrupt the usual way of life leading to loss of life, disability, loss of property, increased financial burdens and psychological stress (Dorfman, 1998). Against this background, diverse methods along the lines of saving, risk sharing and financial pooling mechanisms have been developed to cushion populations against effects of such unavoidable circumstances (Mishr, 2016). Some of the approaches include but are not limited to private savings, provident funds, and insurance (Teyie, 2019). Insurance is one of the regularly used methods to absorb risk and protect life and property, and it involves transferring risks from one individual called insured to another called the insurer (Cummins et al., 2013). Insurance is categorized into life and non-life (or general) insurance, with life insurance involving long-term contracts that provide financial protection in the event of death, and non-life insurance covering short-term risks related to property damage or loss (Raheja, 2017). In life insurance, the policyholder pays regular premiums in exchange for a payout (insured sum) upon death or specified conditions (Ionesco, 2012). However, challenges arise when policyholders fail to pay premiums on time, leading to a loss of coverage and benefits.

Lapsation of life insurance policy is a discontinuation of premium payment by the policyholder during the period of operation of the policy due to any reason other than the death of the policyholder (Vidyavathi, 2013). The policyholder is granted a grace period to clear any unpaid premiums before the lapse occurs, during which the insurer remains obligated to provide benefits in case the insured event takes place. The ability to reinstate a lapsed policy is contingent on the type of policy and can be achieved by fulfilling specific criteria. This therefore affects persistency which is a key performance metric for life insurance companies. Life insurance persistency pertains to the percentage of existing policies from an insurance company that stay active without lapsing or being replaced by policies from different insurers (Teyie, 2019). When calculating the life insurance persistency rate, the emphasis is on comparing the number of active policies to the total policies issued. According to global benchmarks, the minimum persistency rates for policies are set at 90% for the first year, 85% for the second year, and 80% for the third year (AKI, 2017).

Policy lapsation has become a challenge to life insurers globally as it puts pressure on revenue and leads to reduced profits (Vasudev, Bajaj, & Alegre Escolano, 2016). Most insurance companies in Kenya have come up with some form of customer retention programs but these measures have had little impact on the policy lapse issue (Teyie & Justus, 2019). This is because most of these programs have focused more on reactionary rather than proactive measures that address the real drivers of policy lapsation. These include a claw-back program (the commissions paid to the intermediary are recalled by the insurer upon the lapsation of the specific life assurance policy), development of a single premium policies (the insured does not pay periodic premiums but pays one premium to cover the entire insurance period) and the use of automated premium remittance method (Vankayalapati, 2017).

Existing literature on life insurance policy lapses reveals that both policyholder-related and insurer-related factors significantly contribute to lapsation across global markets. Studies from countries like Germany (Dieter & Kiesenbauer, 2011), the USA (Fier et al., 2012; Diulio, 2020), and India (Shodganga, 2012; Subhashini et al., 2016) suggest that macroeconomic conditions,

Stratford Peer Reviewed Journals and Book Publishing
Journal of Information and Technology
Volume 9||Issue 1||Page 105-125 ||December||2025|
Email: info@stratfordjournals.org ISSN: 2617-3573

income shocks, and interest rates influence lapse behavior, particularly in unit-linked products. Factors such as policyholder age, payment mode, premium frequency, employment status, and income levels have been consistently associated with lapse rates. Younger policyholders and those facing financial instability are more prone to lapse, as seen in studies from Nigeria (Mojekwu, 2011), Sri Lanka (Jayetileke et al., 2017), and Kenya (Teyie et al., 2019). Furthermore, annual premium payments and automatic billing tend to reduce lapse rates, while poor persistency is often found among low-income, self-employed, or less educated policyholders.

Policy lapsation is often driven by poor customer service and the actions of insurance agents, including mis-selling and prioritizing commissions over client needs, as seen in countries like India (Anagol et al., 2013) and Zambia (Mtonga, 2021). Additionally, factors such as low awareness, economic strain, and lack of product suitability significantly impact policy persistency and pose financial risks to insurers, particularly in markets like South Africa (KPMG, 2018). Considering that lapsation is a multi-faceted issue, involving economic conditions, customer demographics, intermediary behavior, and institutional practices, risk modeling to predict the probability of policy lapse alongside the major significant factors/ predictors.

Modelling lapses is important to manage and control future risks or uncertainties that may arise in the insurance business. An increase in the lapse rate directly affects the company's book size, product pricing, statutory reserve, trigger insolvency, market-consistent embedded value, and other risk management decisions resulting to lesser new entrants and more policyholders lapsing (Eling & Kochanski, 2012; Barsotti, Milhaud & Salhi, 2016). It is important for life insurers to properly assess and model their exposure to lapse risks and understand cancellations behaviour as accurately as possible (Barsotti et al., 2016).

Currently, the Kenyan Life Insurance market has limited comprehensive lapse model that considers the factors contributing to policy lapses. The models used are based on a simple analysis of previous experience. Lapses in the first year for a new life insurance product are regarded as equal to the lapse rate over the first year for a previous life insurance product. The analysis does not view the lapse rate as a dependent variable whose determination is based on several contributing factors, but rather as a rate dependent only on the type of product in question. In other words, the underlying factors that may contribute to lapse are not viewed individually, neither are they varied or even tested for significance (Ocheche, 2009). Furthermore, traditional statistical models have been applied in such analyses, and the aforementioned modes are usually based on very restrictive assumptions that might be difficult to meet.

Machine learning (ML) models have gained application in many fields and have raised interest in such areas to identify hidden patterns and gain in-depth insights from data. The methods are robust and are capable of handling huge amounts of data marred with complex interrelations. This therefore highlights a limitation in the current models being employed and hence the need to get a better method or model of predicting policy lapse. This paper uses novel machine learning models to unravel patterns and insight on policy lapse rate drivers in the life insurance sector in Kenya. The specific objectives were to determine policy lapsation rate for the different products offered in the industry, compare the performance of supervised machine learning models in predicting policy lapse, identify the key features contributing to policy lapsation in the life insurance sector in Kenya and propose a policy lapse predictive model for external use.

Machine learning is a subset of artificial intelligence (AI) that enables machines to learn and adapt through experience. When implemented effectively, ML can allow organizations to utilize data collection for business benefits (Alzubi et al., 2018). ML techniques analyze potential relationships

Stratford Peer Reviewed Journals and Book Publishing
Journal of Information and Technology
Volume 9||Issue 1||Page 105-125 ||December||2025|
Email: info@stratfordjournals.org ISSN: 2617-3573

between independent and one or multiple dependent variables and ultimately can identify a function that accurately predict a target attribute based on available input attributes (Varian, 2014). There are a wide range of ML algorithms, including Naïve Bayesian Classifiers, Logistic Regression (LR), Decision Trees (DT), K-Nearest-Neighbour (knn), and Neural Networks (NN). ML models have become increasingly important in the context of modelling insurance data, as they simplify various types of data sets (Makariou & Chen, 2021). These models can enhance actuaries' understanding of problems and data, by utilizing unstructured data directly. Although the field of ML is expanding and has great potential for use in actuarial science, it is still recent and not neatly organized (Blier-Wong & Marceau, 2021).

While there exist numerous ML algorithms, each ML tend to work differently for different activities in different environments, contexts and data structure (Kuhn & Johnson, 2013a; Singh et al., 2016). Based on this, the emphasis is on Supervised Models for classification, including Logistic Regression, K-Nearest Neighbours, Neural networks, Tree-Based Approaches such as Regression and Classification Trees, and ensemble models like Bagged Classification Tree, Adaptive Boosting, Extreme Gradient Boosting, and Random Forest. Additionally, the unsupervised algorithm K-means is employed for feature engineering purposes.

2. Methodology

2.1 Data and Variables

This study used secondary data obtained from a life insurance company operating in Kenya. The data covers six years (2018 to 2023). The dependent variable of the study is the lapse which is binary in nature denoted in this study whether a policy is "lapsed" or "not lapsed". The response variable was modelled to identify the independent variables that are statistically significant in explaining future lapse behaviour. This allows for more accurate predictions of future lapse rates and the setting of reliable assumptions for projections. The explanatory variables were the policy features and policyholder features. The policyholder features in the data are policy identifier, entry age, education level, gender, occupation, mode of payment, frequency of payment, marital status, nationality, county of residence and number of dependents. The policy features used in this study include product type, premium amount, sum assured, contract duration, branch, premium frequency, policy term, and payment mode.

2.2 Models

To predict the likelihood of policy lapsation, the study adopted Logistic regression, Neural Network, Random Forest, Extreme Gradient Boost (XGBoost) and Adaptive Boosting (AdaBoost).

2.2.1 Logistic regression

The logistic regression model used is defined as follows.

Let y_i denote the policy lapse status of individual i, y_i is binary in nature with $y_i = 1$ if individual i's policy is lapsed and 0 otherwise.

$$y_i \sim \text{bernoulli}\ (p_i)$$

$$logit(p_i) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \ldots + \beta_p x_p$$

Stratford Peer Reviewed Journals and Book Publishing
Journal of Information and Technology
Volume 9||Issue 1||Page 105-125 ||December||2025|
Email: info@stratfordjournals.org ISSN: 2617-3573

Where,

x_1 ... x_p are features as identified in section 3.1 above. β_1 ... β_p are the model coefficients and β_0 is the intercept. These parameters reflect the association between independent and dependent variables. p_i is the probability of the policies lapsing. Estimation of the model coefficients is usually done using the maximum likelihood approach.

2.2.2 Neural Network

A neural network is usually made of several units called *neurons* of the form

$$h_j(x) = \sigma\left(w_j + \sum_{i=1}^{n} w_{ij} x_i\right)$$

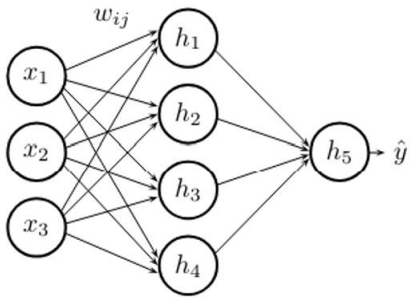

Figure 1: An artificial neural network.

where σ is a non-linear activation function, such as a sign function, the sigmoid function or the softmax activation function. In most cases, these units are organized into successive layers, with the output of one layer transmitted to the next through weighted connections, often referred to as synapses. For example, Figure 1 illustrates a three-layered neural network. The first layer is the input layer, which transmits the input values $x = (x_1, \ldots, x_p)$ identified in section 2.1, to the second layer. The second layer is made of activation units h_j, taking as inputs the weighted values of the input layer and producing non-linear transformations as outputs. The third layer is made of a single activation unit, taking as inputs the weighted outputs of the second layer and producing the predicted value $\hat{y}$. Assuming that this structure is fixed and that all units from the network make use of the same activation function σ, the hypothesis space $\mathcal{H}$ therefore includes all models φ of the form

$$\varphi(x) = \sigma\left(w_5 + \sum_{j=1}^{4} w_{j5}\sigma\left(w_j + \sum_{i=1}^{p} w_{ij} x_i\right)\right)$$

where,

σ is an activation function, a mathematical function which determines the output of a neuron in a neural network, x_i are features as identified in section 3.1 above, and w_{ij} are weights in the neural networks.

Activation functions are specifically used in artificial neural networks to transform an input signal into an output signal, which is then passed as input to the next layer in the network. In an artificial neural network, we first calculate the weighted sum of the inputs and then apply an activation function to this sum to produce the output of the layer. This output is subsequently fed as input to the following layer. A neural network works just like a linear regression model where the predicted output is same as the provided input if an activation function is not defined. The choice of an

Stratford Peer Reviewed Journals and Book Publishing
Journal of Information and Technology
Volume 9||Issue 1||Page 105-125 ||December||2025|
Email: info@stratfordjournals.org ISSN: 2617-3573

activation function is motivated by the fact that a binary class label needs to be predicted (Charu, 2018). The most commonly used activation functions are non-linear functions. One such function is the Sigmoid activation function, which transforms values into the range [0,1]. It can be defined as:

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

The sigmoid function is continuously differentiable and has a smooth S-shaped curve. Its derivative is:

$$\sigma'(x) = \sigma(x)(1 - \sigma(x))$$

However, the sigmoid function is not symmetric around zero, meaning all output values of neurons have the same sign (positive). This lack of symmetry can sometimes slow down learning, but it can be partially addressed by scaling the sigmoid function. Another popular activation function is the Tanh function (Hyperbolic Tangent), which is similar to sigmoid but symmetric about the origin. This symmetry results in outputs ranging between -1 and 1, allowing the neuron outputs to have different signs. It can be defined in terms of the sigmoid function as:

$$\tanh(x) = 2\sigma(2x) - 1$$

Like sigmoid, the tanh function is continuous and differentiable. The third commonly used activation function is the Rectified Linear Unit (ReLU). One advantage of ReLU is that not all neurons are activated simultaneously a neuron is only deactivated when its linear transformation output is zero or negative. Mathematically, ReLU is defined as:

$$f(x) = \max(0, x)$$

(ReLU and other details are discussed in Aggarwal et al. (2018).

In multilayer neural networks, the networks typically follow a feed-forward architecture, where each neuron in a layer connects to every neuron in the next layer, passing data forward through weighted inputs, bias terms, and activation functions. Hidden layers are central to deep learning, acting as the computational core that allows neural networks to approximate functions and detect patterns in data (Charu, 2018).

2.2.3 Random Forest

Random Forest (RF) is a technique used for prediction and behaviour analysis. It is an ensemble method based on bagging and is built upon multiple decision trees (Breiman, 2001). An RF is composed of multiple decision trees that are designed to be uncorrelated with one another. During a classification task, each decision tree independently evaluates the input and produces a classification. The final prediction is typically made by aggregating these individual outputs, often through majority voting. A random forest is a classifier consisting of a collection of tree-structured base classifiers $\{h(x, \Theta_j), j = 1, \ldots.\}$ such that the $\{\Theta_j\}$ are independent identically distributed random vectors where each tree casts a unit vote for the most popular class at input x. Consider training dataset $D = \{(x_1, y_1), \ldots, (x_N, y_N)\}$, where $x_i = (x_{i,1}, \ldots, x_{i,p})^T$ denotes the p predictors and y_i denotes the response, and a particular realization θ_j of Θ_j. The fitted tree is denoted as $\hat{h}_j(x, \theta_j, D)$.

In an RF setup, each decision tree gets a classification result of its own, and the RF takes the classification result which has the highest number of results among all decision trees as the final

Stratford Peer Reviewed Journals and Book Publishing
Journal of Information and Technology
Volume 9||Issue 1||Page 105-125 ||December||2025|
Email: info@stratfordjournals.org ISSN: 2617-3573

result. Each tree in the classification takes input from the samples in the initial data set. Then features are randomly selected for growing the tree at each node. Each tree in the forest should not be pruned until the prediction is reached decisively at the end of the exercise. In this way, the RF enables any classifier with weak associations to create a stronger classifier. It is relatively fast to train, easy to make into a parallel method, simple to implement, and easy to balance the error for unbalanced data sets. However, its classification performance is not good for small sample data and low-dimensional data (with a small number of features) because of the overfitting problem.

2.2.4 Extreme Gradient Boost (XGBoost)

XGBoost is a decision tree-based ensemble algorithm that leverages the concept of gradient boosting to create a strong model. XGBoost works by creating a set of decision trees iteratively, each tree attempting to correct the mistakes of the previous tree. The algorithm employs a gradient descent algorithm to minimize a cost function, which is the sum of the errors of each tree in the ensemble. The final model is a weighted combination of all the decision trees, with each tree assigned a weight based on its contribution to the cost function. Similarly, to gradient boosting, XGBoost builds an additive model by sequentially minimizing loss function, thereby expanding the objective function at each iteration (Singh & Agarwal, 2023).

XGBoost employs a greedy algorithm to construct decision trees by initiating with a single root node and recursively splitting the data into two child nodes. Each split is determined by selecting the feature and threshold that yield the greatest reduction in the loss function. This recursive process continues until a predefined stopping criterion is met, such as reaching the maximum tree depth or the minimum number of samples required in a leaf node. To further mitigate overfitting, XGBoost incorporates a pruning strategy that removes nodes which do not contribute significantly to the loss reduction, thereby simplifying the model without sacrificing predictive performance.

$$L_{xgb} = \sum_{i=1}^{N} L\left(y_i + F(x_i)\right) + \sum_{m=1}^{M} \Omega(h_m)$$

$$L_{xgb} = \sum_{i=1}^{N} L\left(y_i + F(x_i)\right) + \sum_{m=1}^{M} \Omega(h_m)$$

Gradient descent is used in XGBoost to iteratively minimize a loss function by adjusting predictions in the direction of the negative gradient, which indicates the error. XGBoost builds the model by adding decision trees one at a time, with each tree correcting the errors of the previous ones using the computed gradients. To prevent overfitting, XGBoost employs a shrinkage technique (learning rate) that limits each tree's contribution, improving model robustness but requiring more iterations for high accuracy (Singh & Agarwal, 2023).

2.2.5 Adaptive Boosting (AdaBoost)

AdaBoost is an ensemble learning algorithm that combines multiple weak learners to make accurate predictions. It trains a set of weak learners iteratively on a weighted version of the data. Misclassified data points have their weights increased to focus more on them, helping Adaboost learn a set of weak learners that specialize in different parts of the data and work together for accurate prediction (Ying et al., 2013). Adaboost works by combining the predictions of multiple weak learners to make accurate predictions. The final prediction of the Adaboost algorithm is a weighted sum of the predictions of the individual weak learners. The weights of the weak learners

Stratford Peer Reviewed Journals and Book Publishing
Journal of Information and Technology
Volume 9||Issue 1||Page 105-125 ||December||2025|
Email: info@stratfordjournals.org ISSN: 2617-3573

are determined by their accuracy on the training data. The more accurate a weak learner is, the more weight it is given in the final prediction (Ying et al., 2013).

2.3 Data Pre-Processing

This is a crucial step in the ML process as it can significantly influence the performance and accuracy of predictive models. It involves preparing raw data to make it suitable for modeling, which includes tasks such as cleaning the data, handling missing values, transforming variables, and scaling features to a common range. Common pre-processing techniques include data cleaning, feature scaling and centering, encoding categorical variables and feature selection. These steps help in ensuring that the data is of high quality and that models can learn effectively from it.

2.3.1 Data Cleaning

Data cleaning is a crucial step in preparing datasets for analysis, involving the identification and correction or removal of incorrect, noisy, or irrelevant data (Maharana et al., 2022). Data cleaning was done by removing duplicate or irrelevant entries, often caused by merging data from multiple sources; fixing structural errors like typos or mislabeled categories; handling missing values using techniques such as mean, median, or mode imputation; and resolving inconsistent values by ensuring uniform data types.

2.3.2 Feature Scaling and Centering

To undertake feature scaling, data scaling (Peshawa et al., 2014) was used. Data was transformed to take up values between zero and one using the formula:

$$Z = \frac{x - x_{min}}{x_{max} - x_{min}}$$

where Z is new normalized/standardized value, x_i is the data point $(x_1, x_2, \dots, x_n)$, is the sample mean, is the sample standard deviation, x_{min} is the sample minimum and x_{max} is the sample maximum.

2.3.3 Feature Selection

Feature selection was applied to identify the most relevant dependencies or correlations between input features and the target variable. The primary objective was to reduce the number of redundant or irrelevant features, thereby decreasing model complexity and training time, while potentially improving classification accuracy and reducing the risk of overfitting (Haar et al., 2019).

2.3.4 Model Training

Data splitting methodology, tuning and improving machine learning models and k-fold cross validation were used for model training. The train-test split ratio was set to 70:30, where the majority of the data (70%) was allocated to model training and the remaining 30% was used for validation and testing.

To optimize model performance and prevent overfitting or underfitting, hyperparameter tuning was used. Grid search and random search were used for data tuning. Grid search systematically evaluates all possible combinations of hyperparameters, often leading to improved performance but at the cost of high computational time (Hossain & Timmer, 2021). In contrast, random search selects combinations at random, offering a faster alternative and the potential to discover effective or novel parameter sets, though it may not always yield optimal results.

Stratford Peer Reviewed Journals and Book Publishing
Journal of Information and Technology
Volume 9||Issue 1||Page 105-125 ||December||2025|
Email: info@stratfordjournals.org ISSN: 2617-3573

K-fold cross-validation is a robust technique for evaluating the generalization performance of machine learning models by splitting the dataset into *k* equal subsets (folds), training the model on *k–1* folds and validating it on the remaining one, then repeating the process *k* times to average the performance metrics (Santos et al., 2018). Ten-fold stratified cross-validation was used by dividing the data into 10 equal-sized folds, and the samples chosen in a way that each fold contained roughly the same proportions of samples of each target class. While k-fold cross-validation estimates a model's predictive power, bootstrap techniques are used to assess variability in performance estimates, enabling more informed decision-making through confidence intervals.

2.3.5 Model Evaluation

During model evaluation, Confusion Matrix, Receiver Operating Characteristic (ROC) Curve and Area Under Curve (AUC-ROC) and Precision-Recall Curve and Area Under Curve (AUC-PR) were employed.

Confusion matrix

The confusion matrix is a tabular representation of true and predicted class labels, and it is used to evaluate the model's accuracy, precision, recall, and other relevant statistics. In the case of insurance Lapse, a classification model can predict the probability of lapsing, allowing the company to make more informed decisions. The selection of a specific target class was determined by a threshold, typically set at 0.5, and the model's performance is often evaluated through the use of graphical probabilistic performance measures such as ROC curves and precision-recall curves (Kuhn & Johnson, 2013b). From the Confusion Matrix, accuracy was used to measure how the model correctly predicts insurer lapsation and a Kappa statistic was used to measure model's reliability.

Additional metrics employed included prevalence, which measures the proportion of actual positive cases in the dataset, and detection prevalence, which captures the proportion of predicted positives. Positive Predictive Value (PPV) and Negative Predictive Value (NPV) provide the likelihood that positive or negative predictions are correct, incorporating sensitivity, specificity, and prevalence. Lastly, logarithmic loss (log loss) is a more advanced evaluation measure that penalizes incorrect predictions based on the predicted probabilities rather than hard class labels. It is particularly effective for models like logistic regression and neural networks, where probability outputs are critical, and is especially favored in contexts with class imbalance due to its strict penalization of confident but incorrect predictions (Nasteski, 2017).

ROC and AUC

ROC curve was used to evaluate the relationship between True Positive (TP) rate and False Positive (FP) rate. The ROC plots the trade-off between the true positive rate and the false positive rate at different thresholds. The ROC curve can be used to determine alternate cut-off values for class probabilities. The ROC can also be measured as a single metric by calculating the AUC-ROC. The AUC of the ROC curve summarizes its overall performance, with a score of 1 representing a perfect classifier and 0 indicating a poor classifier (Kuhn & Johnson, 2013b). AUC ROC, which considers both true positive and false positive rates, is a preferred metric in these cases by providing a more comprehensive evaluation of classifier performance, especially when the positive class is rare or the class distribution is imbalanced (Agarwal et al., 2016).

Stratford Peer Reviewed Journals and Book Publishing
Journal of Information and Technology
Volume 9||Issue 1||Page 105-125 ||December||2025|
Email: info@stratfordjournals.org ISSN: 2617-3573

Precision-Recall Curve and Area Under Curve (AUC-PR)

Precision-Recall curves is used to evaluate the performance of a classification model in identifying a particular type of data by plotting precision on the y-axis and recall on the x-axis, and a model with high scores in both is considered effective. The area under the Precision-Recall curve, obtained by integrating precision and recall values at various thresholds, is another evaluation metric for binary classifiers. The higher the area under the curve, the better the classifier's performance (Boyd et al., 2013). In this study, the algorithms were evaluated using AUC-ROC, sensitivity and NPV.

2.3.6 Model Selection

The evaluation of models was performed using a variety of metrics such as accuracy, precision, recall, F1 score, Logarithmic Loss, among others. Cross validation is a widely used technique for estimating the performance of a model on unseen data and involves dividing the data into multiple subsets, training the model on one subset, and evaluating its performance on the remaining subsets. The average performance across all subsets is then used as an estimate of the model's performance on unseen data. In model selection, the best model was typically chosen based on its predictive accuracy, as predicted by AUC-ROC. These metrics provide a quantitative way to compare the performance of different models.

3. Model Setup

3.1 Data Preparation

Secondary data for a life insurance company operating in Kenya was collected covering the period 2018 to 2023 with policy and policyholder features. The dataset initially consisted of 11 categorical variables and 7 numerical variables with 21,884 unique policyholders who are principal members. Three files were extracted i.e., policyholder data, policy data and dependents data. Policyholder data had ten variables namely, policy number which is the policy identifier, age, education level, gender, occupation name, mode of payment, frequency, marital status, nationality and town. Policy data contained policy number, effective date, product type, premium amount, sum assured, contract duration/term, agent number code, branch code, premium frequency, term of policy, payment mode code and policy status. The dependents data contained policy numbers and the number of dependents.

The three files were merged using the policy number as the unique identifier and any duplicated variables in the policy and policyholder data were deleted. Policy status, the outcome variable had two categories "lapsed" or "not lapsed". Four policies were dropped since they missed the gender, and this is not easily derived from any other variable in the dataset. The dataset remained with 18,680 policyholders.

In addressing missing data, the education level variable was excluded from analysis due to the complete absence of observations (100% missing). For variables with substantial missing data proportions - marital status (74% missing) and nationality (89% missing) - appropriate imputation methods were implemented following established statistical practices (Schafer & Graham, 2002). Marital status was completed through logical imputation using the dependents variable as a proxy indicator. Here, policyholders reporting no dependents were classified as single and those with dependents categorized as married. This approach reflects common techniques for handling missing categorical data through logical relationships between variables (van Buuren, 2018).

Stratford Peer Reviewed Journals and Book Publishing
Journal of Information and Technology
Volume 9||Issue 1||Page 105-125 ||December||2025|
Email: info@stratfordjournals.org ISSN: 2617-3573

The nationality variable was imputed using the modal value (Kenyan), a decision supported by both the insurance company's domestic operations and branch location data. This method aligns with recommendations for imputing highly skewed categorical variables where one category dominates the distribution (Allison, 2001). While such single imputation methods provide practical solutions for analysis, researchers acknowledge they may underestimate variance in the dataset (Rubin, 1996).

Further, certain machine learning libraries cannot process categorical variables directly. Thus, some categorical variables were transformed into numerical values using one-hot encoding, a technique particularly effective for nominal categories. These include low-cardinality categorical variables i.e., `Premium Frequency`, `Payment mode description`, `Gender`, `Mode of payment`, `Marital Status` and `Nationality`. This method involved first mapping categorical values to integers (ranging from 0 to n-1, where n is the number of unique categories, assigned alphabetically) and then converting these integers into binary indicators (0 or 1). The high-cardinality categorical variables like `Product Type`, `Branch`, `Occupation name`, `Town` were encoded using Target Encoding. This method replaces each category with the mean of the target variable (policy status) for that category, capturing the relationship between the categorical feature and the target.

During Feature Scaling, numerical variables such as 'Premium Amount,' 'Sum Assured,' and 'Age' were normalized using StandardScaler. This standardization process adjusts the features to a common scale, preventing variables with larger magnitudes from disproportionately influencing the model while ensuring balanced contributions from all features during training. Finally, features such as `Effective_date` were dropped, as they were not relevant for prediction. The remaining features were used as input variables for the models. The target variable “Policy status” is categorical with two levels, 1 if policy lapsed and 0 if policy in force.

3.2 Model Selection and Training

Five machine learning models were selected for training and evaluation: Logistic Regression, ANN, Random Forests, XGBoost and AdaBoost. Each model was trained on the standardized training set using default hyperparameters. For Logistic Regression, the maximum number of iterations (`max_iter`) was increased to 1000 to ensure convergence. A comparison of the training and testing data was done to show the characteristics of the various variables across the two data sets. This is shown in Table 1.

Stratford Peer Reviewed Journals and Book Publishing
Journal of Information and Technology
Volume 9||Issue 1||Page 105-125 ||December||2025|
Email: info@stratfordjournals.org ISSN: 2617-3573

Table 1: *Distribution of Study Variables for Policy Data by Training and Testing Data Sets (2018-2023)*

Study variables	Training Data N (%)	Testing Data N (%)	P-value
Gender (n %)			
- Female	5 544 (42.4)	2 329 (41.6)	0.2133
- Male	7 533 (57.6)	3 274 (58.4)	
Product type(n %)			
- Hospital Cash Plan	535 (4.1)	226 (4.0)	0.2543
- Critical Care	4058 (31.0)	1747 (31.2)	
- Dahari Dumu	50 (0.4)	11 (0.2)	
- Investment	564 (4.3)	225 (4.0)	
- Select	18 (0.1)	6 (0.1)	
- Super 7	3848 (29.4)	1705 (30.4)	
- Term Life A	1779 (13.6)	759 (13.5)	
- Term Life B	33 (0.3)	18 (0.3)	
- Life Plan A	9 (0.1)	3 (0.1)	
- Life Plan A Limited To 60 Years	466 (3.6)	191 (3.4)	
- Life Plan A Limited To 65 Years	32 (0.2)	12 (0.2)	
- Life Plan B	38 (0.3)	14(0.2)	
- Life Plan B Limited To 60 Years	168 (1.3)	71 (1.3)	
- Life Plan B Limited To 65 Years	12 (0.1)	4 (0.1)	
- Memorial	25 (0.2)	8 (0.1)	
- Income Protection	10 (0.1)	4 (0.1)	
- Retrenchment	285 (2.2)	117 (2.1)	
- Last Expense	1147 (8.8)	482 (8.6)	
Premium frequency			
- Half Yearly	190 (1.5)	85 (1.5)	0.2289
- Monthly	10425 (79.7)	4487 (80.1)	
- Quarterly	523 (4.0)	221 (3.9)	
- Yearly	1939 (14.8)	810 (14.5)	
Payment mode description			
- Cash	670 (5.1)	302 (5.4)	0.234
- Checkoff	3972 (30.4)	1744 (31.1)	
- Cheque	1393 (10.7)	599 (10.7)	
- Direct Debit	3710 (28.4)	1565 (27.9)	
- Mobile Money	3267 (25.0)	1373 (24.5)	

Stratford Peer Reviewed Journals and Book Publishing
Journal of Information and Technology
Volume 9||Issue 1||Page 105-125 ||December||2025|
Email: info@stratfordjournals.org ISSN: 2617-3573

- Standing Order		55 (0.4)	19 (0.3)	
- Through Eft		10 (0.1)	1 (0.0)	
Age, mean (SD)		33.6 (13.7)	33.8 (13.7)	<0.001
Premium amount, mean (SD)		7 397.3 (43 775.8)	7 521.1 (35 528.5)	<0.001
Term of policy, mean (SD)		32.1 (24.4)	31.8 (24.3)	<0.001
Sum assured, mean (SD)		1 039 743.3 (4 882 984.2)	1 057 544.3 (5 136 051.3)	<0.001
Marital status (n %)				
- Divorced		7 (0.1)	2 (0.0)	0.2426
- Married		12128 (92.7)	5194 (92.7)	
- Other		30 (0.2)	10 (0.2)	
- Single		912 (7.0)	397 (7.1)	
Branch				
- Corporate business		6 (0.0	5 (0.1)	0.0923
- Direct h/o		303 (2.3)	145 (2.6)	
- Eldoret		2234 (17.1)	983 (17.5)	
- Hnw		1019 (7.8)	439 (7.8)	
- Kisumu		1842 (14.1)	719 (12.8)	
- Metropolitan nrb		6 (0.0)	2 (0.0)	
- Mombasa		2006 (15.3)	868 (15.5)	
- Nakuru		525 (4.0)	204 (3.6)	
- Ifa		9 (0.1)	3 (0.1)	
- Platinum nrb		1382 (10.6)	630 (11.2)	
- Premier nrb		3709 (28.4)	1016 (28.4)	
- Prestige nrb		36 (0.3)	14 (0.2)	
Nationality				
- Foreigner		9(0.1)	3(0.1)	0.2133
- Kenyan		13 068 (99.9)	5 600 (99.9)	
Town	Yes		Yes	<0.001
Occupation	Yes		Yes	<0.001
Dependants	Yes		Yes	0.171

Stratford Peer Reviewed Journals and Book Publishing
Journal of Information and Technology
Volume 9||Issue 1||Page 105-125 ||December||2025|
Email: info@stratfordjournals.org ISSN: 2617-3573

4. Results

4.1 Data Exploration

Table 2 shows that policy lapses are influenced by various factors. Demographically, there are more male policyholders (58%) compared to female policyholders (42%). The males exhibit a higher lapse rate (62.9%) compared to females (47.8%). The average policy entry age for lapsed policyholders is 30.5 years, compared to 37.8 years for retained policyholders, indicating a higher lapse tendency among younger policyholders.

Premium payment patterns i.e., the frequency and mode of payment also affect policy lapsation. The yearly premium payments (which entail a large lump sum amount paid at once) have the highest lapse rate (72.9%). On the other hand, monthly payments have a lower lapse rate (53.4%). These are regular smaller payments that are more sustainable. Payment mode appears to be associated with policy lapse rates. Policies paid by cash and standing orders exhibit relatively low lapse rates of 10.7% and 12.2% respectively. In contrast, cheque and mobile money payment modes demonstrate significantly higher lapse rates of 82.6% and 63.2% respectively.

Financial and policy characteristics also exhibit notable differences between lapsed and retained policies. On average, lapsed policies have significantly lower premium amounts (Kshs. 3,535) compared to retained policies (Kshs. 12,504). Similarly, the average sum assured for lapsed policies is substantially lower (Kshs. 473,360) than that of retained policies (Kshs. 1,788,441). Interestingly, lapsed policies tend to have longer policy terms, averaging 34.6 years, compared to 28.7 years for retained policies. This may suggest that longer-term policy commitments are more challenging to maintain over time. Married policyholders dominate with 92.5% lapse rate and 92.9% non-lapse rate. HNW branch has a higher lapse rate (72.6%) while Mombasa has a very low lapse rate (23.2%).

Stratford Peer Reviewed Journals and Book Publishing
Journal of Information and Technology
Volume 9||Issue 1||Page 105-125 ||December||2025|
Email: info@stratfordjournals.org ISSN: 2617-3573

Table 2: *Distribution of Study Variables by Lapse Status (Lapse or No-Lapse, 2018-2023*

Study variables	Lapse N (%)	No-Lapse (In-Force) N (%)	P-value
Gender (n %)			
- Female	3764 (47.8)	4109 (52.2)	<0.001
- Male	6795 (62.9)	4012 (37.1)	
Product type(n %)			
- Hospital Cash Plan	441(58)	320 (42)	
- Critical Care	23(37.7)	38 (62.3)	
- Dahari Dumu	235 (29.8)	554 (70.2)	
- Investment	21(87.5)	3 (12.5)	
- Select	2604 (46.9)	2949 (53.1)	
- Super 7	1477 (58.2)	1061 (41.8)	
- Term Life A	26 (51)	25 (49)	
- Term Life B	0 (0)	12 (100)	
- Life Plan A	154 (23.4)	503 (76.6)	<0.001
- Life Plan A Limited To 60 Years	9 (20.5)	35 (79.5)	
- Life Plan A Limited To 65 Years	10 (19.2)	42 (80.8)	
- Life Plan B	59 (24.7)	180 (75.3)	
- Life Plan B Limited To 60 Years	5 (31.3)	11 (68.8)	
- Life Plan B Limited To 65 Years	2 (6.1)	31(93.9)	
- Memorial	3983 (68.6)	1822 (31.4)	
- Income Protection	8 (57.1)	6 (42.9)	
- Retrenchment	2 (6)	378 (94)	
- Last Expense	1478 (90.7)	151 (9.3)	
Premium frequency			
- Half Yearly	146 (53.1)	129 (46.9)	<0.001
- Monthly	7966 (53.4)	6946 (46.6)	
- Quarterly	444 (59.7)	300 (40.3)	
- Yearly	2003 (72.9)	746 (27.1)	
Payment mode description			
- Cash	104 (10.7)	868 (89.3)	<0.001
- Checkoff	1898 (33.2)	3818 (66.8)	
- Cheque	1646 (82.6)	346 (17.4)	
- Direct Debit	3201 (60.7)	2074 (39.3)	
- Mobile Money	2934 (63.2)	1706 (36.8)	
- Standing Order	9 (12.2)	65 (87.8)	
- Through Eft	3 (27.3)	8 (72.7)	

Stratford Peer Reviewed Journals and Book Publishing
Journal of Information and Technology
Volume 9||Issue 1||Page 105-125 ||December||2025|
Email: info@stratfordjournals.org ISSN: 2617-3573

Age, mean (SD)	30.5 (5.2)	37.8 (9.91)	<0.001
Premium amount, mean (SD)	3535.2 (9594.8)	12 504.3 (61575.4)	<0.001
Term of policy, mean (SD)	34.6 (25.2)	28.7 (22.8)	<0.001
Sum assured, mean (SD)	473,360.5 (1821871.6)	1 788 441.2 (7 162 496.4)	<0.001
Marital status (n %)			
Divorced	3 (33.3)	6 (66.7)	0.1174
Married	9813 (56.7)	7509 (43.3)	
Other	17 (42.5)	23 (57.5)	
Single	726 (55.5)	583 (44.5)	
Branch			
- Corporate business	2 (18.2)	9 (81.8)	<0.001
- Direct h/o	46 (10.3)	402 (89.7)	
- Eldoret	1 513 (47)	1 704 (53)	
- Hnw	1 058 (72.6)	400 (27.4)	
- Kisumu	1 240 (48.4)	1 321(51.6)	
- Metropolitan Nairobi	8 (100)	0 (0)	
- Mombasa	2 206 (23.2)	668 (76.8)	
- Nakuru	356 (48.8)	373 (51.2)	
- Ifa	2 (16.7)	10 (83.3)	
- Platinum Nairobi	919 (45.7)	1 093 (54.3)	
- Premier Nairobi	2 974 (56.1)	2 326 (43.9)	
- Prestige Nairobi	44 (88)	6 (12)	
Nationality			
- Foreigner	0 (0.0)	12 (100)	0.0002522
- Kenyan	10 559 (56.6)	8 109(43.4)	
Town	Yes	Yes	<0.001
Occupation	Yes	Yes	<0.001
Dependants	Yes	Yes	<0.001

https://doi.org/10.53819/81018102t2535

Stratford Peer Reviewed Journals and Book Publishing
Journal of Information and Technology
Volume 9||Issue 1||Page 105-125 ||December||2025|
Email: info@stratfordjournals.org ISSN: 2617-3573

4.2 Model Performance Evaluation and Comparison

The study employed five distinct classification algorithms Logistic Regression, ANN, Random Forest, XGBoost and AdaBoost (AB) to assess their predictive capabilities. Each model was rigorously evaluated using multiple performance metrics, including ROC-AUC, precision-recall AUC (PR-AUC), sensitivity, specificity, accuracy, precision, NPV and Mathew's correlation coefficient (MCC). The dominating best modelling method across the measures was crowned the best modelling method when predicting lapses. Table 3 presents comprehensive performance metrics across all models for the training and testing data sets.

Table 3: *Performance Metrics of Machine Learning Models*

Model	Dataset	ROC-AUC	PR-AUC	Sensitivity	Specificity	MCC	Precision	NPV	Accuracy
LR	Training	0.843	0.876	0.804	0.724	0.529	0.790	0.741	0.769
	Testing	0.841	0.876	0.792	0.737	0.528	0.799	0.728	0.768
ANN	Training	0.994	0.996	0.933	0.980	0.908	0.984	0.919	0.954
	Testing	0.808	0.851	0.724	0.749	0.469	0.792	0.672	0.735
RF	Training	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Testing	0.882	0.912	0.811	0.801	0.608	0.843	0.762	0.806
XGB	Training	0.982	0.986	0.936	0.923	0.858	0.940	0.918	0.930
	Testing	0.875	0.907	0.808	0.797	0.601	0.840	0.758	0.803
AB	Training	0.842	0.881	0.787	0.747	0.532	0.801	0.731	0.769
	Testing	0.837	0.876	0.770	0.758	0.524	0.808	0.713	0.765

On the training data, Random Forest (RF) model emerged as the best performing classifier, achieving a perfect value of 1.0 for all evaluation metrics. The ANN followed closely, demonstrating near-perfect performance with a ROC-AUC of 0.994, PR-AUC of 0.996, accuracy of 0.954, and precision of 0.984, indicating strong learning capability. The XGBoost (XGB) model also performed well, though slightly below RF and ANN, with a ROC-AUC of 0.982, PR-AUC of 0.980, accuracy of 0.930, and precision of 0.840. In contrast, Logistic regression (LR) and AdaBoost (AB) showed moderate performance with ROC-AUC scores of 0.843 and 0.842, PR-AUC of 0.897 and 0.881, accuracy of 0.769 and 0.769 and precision of 0.790 and 0.801 respectively.

RF and XGB emerged as the top-performing models, achieving accuracy scores of 80.6% and 80.3%, respectively. These models also demonstrated strong discriminatory power, with ROC-AUC scores of 0.882 and 0.875 respectively, thus indicating their ability to effectively distinguish between lapsed and in-force policies. Their high precision-recall AUC (PR-AUC) scores of 0.912 and 0.907 further underscored their robustness, particularly in handling imbalanced datasets.

Stratford Peer Reviewed Journals and Book Publishing
Journal of Information and Technology
Volume 9||Issue 1||Page 105-125 ||December||2025|
Email: info@stratfordjournals.org ISSN: 2617-3573

LR and AB displayed moderate performance, with accuracy rates of 76.8% and 76.5%, respectively. While Logistic Regression maintained balanced sensitivity (0.792) and specificity (0.737), AB showed slightly lower sensitivity (0.770) but comparable specificity (0.758). These models, though less accurate than ensemble methods, remain valuable for scenarios where interpretability is prioritized over predictive power.

ANN displayed signs of significant overfitting, achieving near-perfect performance on the training set (accuracy: 95.4%, ROC-AUC: 0.994) but a marked decline on the testing set (accuracy: 73.5%, ROC-AUC: 0.808). This performance gap suggests that the model may be memorizing training data rather than generalizing effectively, indicating the potential need for a larger training dataset and/or the application of more advanced regularization techniques to improve generalization.

The superior performance of ensemble methods like RF and XGB can be attributed to their ability to capture non-linear relationships and feature interactions. These models leverage multiple decision trees to reduce bias and variance, making them particularly effective for complex classification tasks. In contrast, simpler models like LR, while useful for baseline comparisons, lack the flexibility to model intricate patterns in the data. XGB and RF's consistent outperformance against the single classifiers reiterates that ensemble models generally perform better than the single classifiers (Dietterich, 2000; Lessmann et al., 2015; Loisel et al., 2021).

4.3 Feature Importance Analysis

An examination of feature importance scores revealed that policyholder occupation, payment mode, and sum assured were the most influential predictors of lapsation. Occupation type consistently ranked highly across all models, reflecting the impact of employment stability on policy retention. Payment modes such as mobile money and cheques were associated with higher lapse rates, likely due to their irregular nature compared to automated methods like standing orders.

Financial factors, particularly the sum assured and premium amounts, played a critical role in policy lapse behavior. Policies with lower sums assured and premiums were associated with higher lapse rates, indicating that affordability and perceived value may significantly influence policyholder retention. Furthermore, product type emerged as a key determinant of lapse propensity, with certain insurance products exhibiting a greater likelihood of lapsation than others.

These findings align with existing literature, which highlights the interplay between demographic, financial, and behavioral factors in driving policy lapses. For instance, younger policyholders and males were more likely to lapse, possibly due to lower financial literacy or changing priorities. Mojekwu (2011) also showed that in Nigeria, young people take up policies and terminate them early. The reason young people are terminating their policies early is illustrated by Valdez et al. (2014), where he showed that this may be due to the fact that the younger policyholders might still be looking elsewhere. The results underscore the need for insurers to tailor retention strategies based on these insights.

Stratford Peer Reviewed Journals and Book Publishing
Journal of Information and Technology
Volume 9||Issue 1||Page 105-125 ||December||2025|
Email: info@stratfordjournals.org ISSN: 2617-3573

4.4 Policy Lapse Rate per Product

The rate of lapse varies significantly across the different product types as shown in Figure 2.

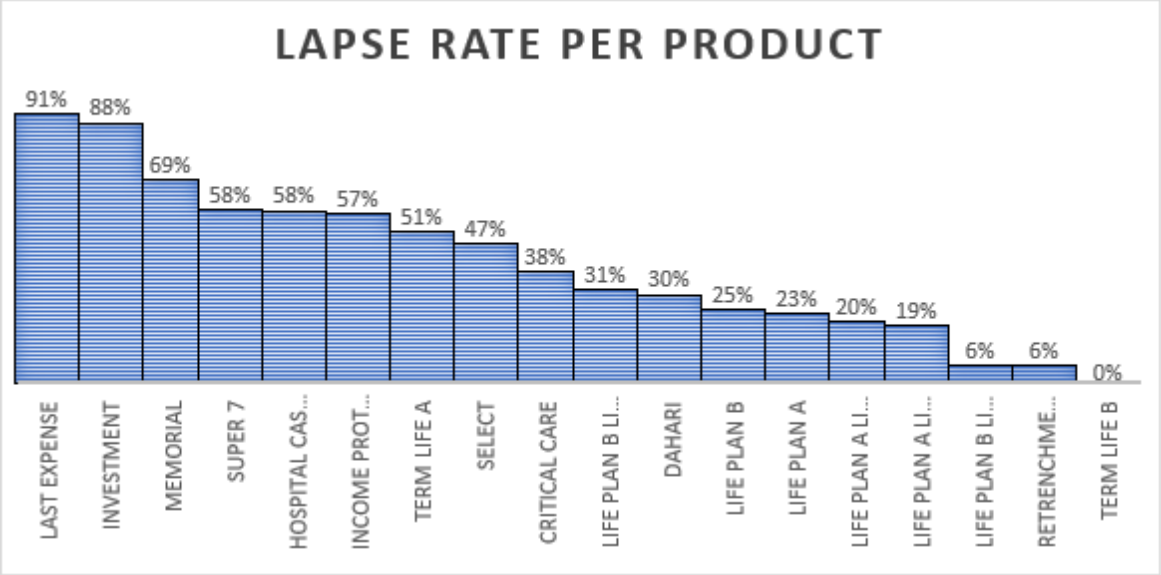

Figure 2: Lapse Rate per Product

Products with higher lapse rates are mostly providing protection against a specific risk as opposed to savings products which combine insurance coverage with a savings or investment component. This is aimed at helping policyholders to build wealth over time while providing some level of protection. This may indicate issues such as affordability, lack of customer engagement, or inadequate product features.

5. Discussions and Conclusions

Excessive policy cancellations pose a significant threat to an insurer's financial stability and public standing. By accurately forecasting potential lapses, companies can develop targeted retention approaches that effectively mitigate customer attrition risks. This article intended to illustrate the predictive power of different machine learning models, their robustness, flexibility, sensitivity and generalization when exposed to a different dataset. It also sought to illustrate the features that drive the lapse rate.

The study revealed that demographic, financial, and behavioral factors significantly influence policy lapsation. Younger policyholders, males, and those with lower sums assured exhibited higher lapse rates. Payment modes such as mobile money and cheques were associated with increased lapsation, while automated methods like standing orders showed better retention.

To predict lapses, insurers could deploy Random Forests or XGBoost. These models enable early identification of at-risk policyholders, allowing for timely interventions. Additionally, insurers should consider tailoring products and payment options to meet the needs of high-risk groups, such as offering flexible premium schedules or financial education programs.

Future research could explore the integration of external datasets, such as macroeconomic indicators, to enhance model robustness and generalizability of predictive models. Additionally, the use of advanced techniques such as hyperparameter optimization and deep learning holds potential for improving model performance and predictive accuracy.

However, the findings of this study are subject to several limitations. The use of data from a single insurer may limit the generalizability of the results across the broader insurance industry. Moreover, the absence of important variables, such as policyholder income, constrained the scope of the analysis. A significant proportion of the dataset contained missing values, necessitating assumptions and imputations that may have introduced bias. Lastly, although strategies were employed to address class imbalance, its residual effects may still influence the model's effectiveness and reliability.

Future research should address these gaps by incorporating broader datasets and exploring hybrid modeling approaches. Real-world pilot testing of the proposed models would also validate their practical utility. By addressing these challenges, subsequent studies can further advance the application of machine learning in insurance risk management.

Conflict of Interest

There is no conflict of interest

References

1. Agarwal, A., C. Baechle, R. S. Behara, & V. Rao (2016). "Multi method approach to wellness predictive modeling." In: *Journal of Big Data*.
2. Aggarwal, C. C. (2018). Neural networks and deep learning (Vol. 10, No. 978, p. 3). Cham: springer.
3. Allison, P. D. (2001). Missing data. Sage Publications.
4. Anagol, S., Cole, S. & Sarkar, S. (2013). *Understanding the Advice of Commissions-Motivated Agents: Evidence from the Indian Life Insurance Market*
5. Barsotti, F., X. Milhaud, & Y. Salhi (2016). "Lapse risk in life insurance: correlation and contagion effects among policyholders' behaviors." *In: Insurance: Mathematics and Economics 71.*
6. Breiman, L. (2001). "Random Forests." *In: Machine Learning* 45(1), pp. 5–32.
7. Charu C. A. (2018). Neural networks and deep learning: a textbook. Springer.
8. Cummins, J. D., Smith, B. D., Vance, R. N., & Vanderhel, J. L. (Eds.). (2013). *Risk classification in life insurance (Vol. 1)*. Springer Science & Business Media.
9. Dietterich, T.G. (2000). Ensemble methods in machine learning. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 1857: 1–15.
10. Dorfman, M. S. (1998). *Introduction to risk management and insurance* (6th ed.). Upper Saddle River, NJ: Prentice Hall.
11. Ionesco (2012). *Life insurance-their characteristics importance and actuality on the Romanian Market.*
12. Kiesenbauer, D. (2012). Main determinants of lapse in the German life insurance industry. *North American Actuarial Journal,* 16(1), 52-73.
13. Kuhn, M. and K. Johnson (2013b). *Applied Predictive Modeling*. Springer Science + Business Media New York.
14. Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling* (Vol. 26, p. 13). New York: Springer.

15. Lessmann, S., Baesens, B., Seow, H.-V. & Thomas, L.C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, 247(1): 124–136.
16. Loisel, S., P. Piette, & C.-H. J. Tsai (2021). "Applying economic measures to lapse risk management with machine learning approaches." In: *ASTIN Bulletin: The Journal of the IAA 51(3),* pp. 839–871.
17. Mishr, K. (2016). *Fundamentals of life insurance theories and applications.* PHI Learning Pvt. Ltd.
18. Mojekwu, J.N. (2011). Study of modes of exit of life-insurance policyholders in Nigeria: Trends and patterns. *International Business Research*, 4(3).
19. Mtonga, W. (2021). *Factors that lead to life insurance policy lapses at zsic life insurance limited* (Doctoral dissertation, The University of Zambia).
20. Ocheche, J. (2009). *Modeling lapse rates using economic variables. A case study for a life insurance company operating in Kenya* (Doctoral dissertation, The University of Nairobi).
21. Peshawa J. Muhammad Ali, & Rezhna H. Faraj (2014). Data Normalization and Standardization: A Technical Report, Machine Learning Technical Reports, 1(1), 1-6.
22. Raheja Bajaj, M. V. (2017). On the Drivers of Lapse Rates in Life Insurance. Barcelona: University of Barcelona
23. Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association,* 91(434), 473-489.
24. Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. Psychological Methods, 7(2), 147-177.
25. Singh, D & Agarwal, S. (2023). *XGBoost & AdaBoost* National Institute of Science Education and Research (NISER), Bhubaneswar
26. Still, L., & Stokes, G., 2016. *Short Term Insurance in South Africa 2016/17*. S and S Analytica.
27. Teyie, S. E., & Justus, T. A. R. I. (2019). Intermediary Factors Affecting Persistency of Ordinary Life Assurance Policies in Kenya. *International Journal of Social Sciences Management and Entrepreneurship (IJSSME)*, *3*(2).
28. Teyie, S. E., & Justus, T. A. R. I. (2019). Intermediary Factors Affecting Persistency of Ordinary Life Assurance Policies In Kenya. *International Journal of Social Sciences Management and Entrepreneurship (IJSSME)*, *3*(2).
29. Van Buuren, S. (2018). Flexible imputation of missing data (2nd ed.). CRC Press.
30. Vankayalapati, P., (2017). *Impact of the lapsation of life insurance policies*. The Chartered Insurance Institute. London.
31. Varian, H. R. (2014). *"Big Data: New Tricks for Econometrics."* In: Journal of Economic Perspectives.
32. Vasudev, M., Bajaj, R., & Alegre Escolano, A. (2016). On the drivers of lapse rates in life insurance. *The Geneva Papers on Risk and Insurance – Issues and Practice, 41*(2), 337–357. https://doi.org/10.1057/gpp.2015.29
33. Vidyavathi, K. (2018) Cost of Lapsation to Policyholders in Indian Life Insurance Industry EPRA *International Journal of Economic and Business Review Vol. 6 (3) 24-29.*
34. Ying, C., Qi-Guang, M., Jia-Chen, L., & Lin, G. (2013). Advance and prospects of AdaBoost algorithm. *Acta Automatica Sinica,* 39(6), 745-758.